Table of Contents

[Deliverables 1](#_Toc30676801)

[Bonus 2](#_Toc30676802)

[Report Introduction (1 mark) 2](#_Toc30676803)

[Exploratory Data Analysis (15 marks) 2](#_Toc30676804)

[Imputing (10 marks) 3](#_Toc30676805)

[Dummy Variables (10 marks) 3](#_Toc30676806)

[Binning (5 marks) 3](#_Toc30676807)

[Modelling (25 marks) 3](#_Toc30676808)

[Model Evaluation (25 marks) 3](#_Toc30676809)

[Model Selection (5 marks) 3](#_Toc30676810)

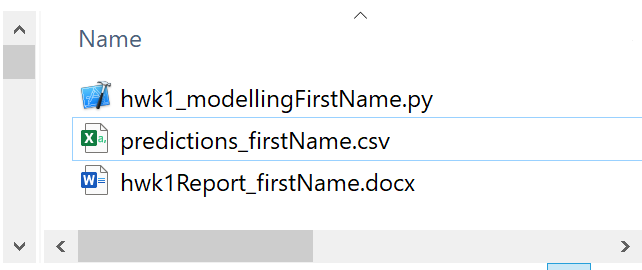
[Model Interpretation (4 marks) 3](#_Toc30676811)

For this homework you will create multiple models and select the best one for predicting housing prices (**SalePrice**). Please use the **hw1\_housingV2.csv** file from the Homework1 folder to train and test your model.

10% will be deducted from the assignment for each day that it is late.

## Deliverables

* In your **zipped** deliverable folder please submit only three files:



* + \*one\* code file named **hwk1\_modellingFirstName.py**. This file must contain all code to allow me to run the project from start to finish without error. The code should perform all exploratory data analysis, data preparation, modeling, model validation and scoring to output the predictions for the mystery file input to a csv file. This code must include the four models which are included with your model summary in your report. One of the four models is **your** selected model. At the very start of this file include your file path declarations as follows so I can easily modify the code to make it work on my machine. Your code file declares the following file paths so I can test your code easily:

PATH = "/Users/pm/Desktop/DayDocs/"

IN\_FILE = "hw1\_housingV2.csv" # Do not change.

FULL\_PATH = PATH + IN\_FILE # Do not change.

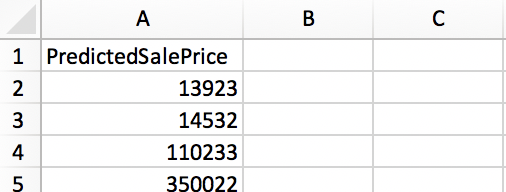
MYSTERY\_FILE = "hw1\_housing\_mysteryfileV2.csv" # Do not change.

OUT\_FILE = "predictions\_yourFirstName.csv" # Do not change.

IN\_PATH = PATH + MYSTERY\_FILE # Do not change.

OUT\_PATH = PATH + OUT\_FILE # Do not change.

* + Your report (MAXIMUM 4-8 Pages) – Be creative here. Initiative, good self-judgment and professional delivery are appreciated.
  + \*one\* csv file named **predictions\_firstName.csv** which is the output from running the prediction code in hwk1\_modelFirstName.py. If needed please see <https://datatofish.com/export-dataframe-to-csv/>
  + This **predictions\_firstName.csv** file must contain only the content as shown (one column only with matching header) in this exact format:



I require this format to run it through my own validator to calculate the RMSE.

Marks will be deducted if deliverables are not submitted in the format requested.

## Bonus

* The person in the class who gets the lowest RMSE when I test their predictions code without error will get a whopping 2% added to their overall course mark.
* The next two lowest RMSE’s from running I test their predictions code without error will get a 1% bonus added to their overall course mark plus a free pass to show boat if they choose in front of the class for 15 seconds.
* All participants who make an honest effort also earn glory that lasts forever.

## Report Introduction (1 mark)

* Describe the problem that you are attempting to solve.

## Exploratory Data Analysis (15 marks)

For the EDA section please assume you are presenting an overview of the data to a not-very-technical group of managers. Please use good sense when assembling the EDA. Be mindful of your time and also of the time for the group you are serving.

* Prepare a summary of the data.
* Show the correlations.
* Highlight columns of interest and how they might impact the predictions positively or negatively.
* You may show scatter plots, histograms, bivariate plots where appropriate.

## Imputing (10 marks)

* Impute values for OverallQual, YearBuilt, GrLivArea. Create columns with the names m\_OverallQual, imp\_OverallQual, m\_YearBuilt, imp\_YearBuilt, m\_GrLivArea, imp\_GrLivArea. You are welcome to choose any technique desired for imputing as long as you implement it in your code.

## Dummy Variables (10 marks)

* Create dummy variables for at least two of the non-numeric variables.

## Binning (5 marks)

* Create categories to group some at least one column of numeric data.

## Modelling (25 marks)

* For this assignment please use an 80/20 split for train/test.
* Be sure to include m\_OverallQual, imp\_OverallQual, m\_YearBuilt, imp\_YearBuilt, m\_GrLivArea, imp\_GrLivArea at least once in your models.
* Be sure to include dummy variables at least once in your models.
* Be sure to include binned columns at least once in your models.

## Model Evaluation (25 marks)

In your report:

* Compare the results of your best four models in a table with any scores that you find helpful. Be sure that at least one of these models includes columns for all imputed, dummy and binned columns.
* The table must list all columns included in the model along with scores that include scores for RMSE, , , AIC, BIC.
* For your selected model, show a plot of the predicted vs. actual results, residual error versus actual results, a histogram of the residuals and a qq plot of the residuals.

## Model Selection (5 marks)

In your report identify your selected model and explain why.

## Model Interpretation (4 marks)

In your report:

* Write out the equation for your model.
* Include a descriptive explanation in English about your model which discusses how the variables interact with each other.

Tip 1: Since the data set is so wide you will find these commands will help to ensure the output is more reader-friendly in your code console:

* print(df.head().transpose())
* print(df.describe().transpose())

Tip 2:

**Strongly Recommended:** Use methods to modularize and re-use your code. Get comfortable running selected portions of the code only rather than all of your Python code at once. I have had trouble with this feature in PyCharm but it does work in Spyder. In any case, choose an editor that works best for you.